**ATTACHMENT B:**

EG&G STATISTICAL APPLICATIONS GROUP [SA] COMMENTS ON:

# STATISTICAL METHODOLOGY FOR BACKGROUND AND COMPARISONS AT THE ROCKY FLATS PLANT

PREPARED BY

DR. R. O. GILBERT.

# REVIEW OF DR. R. O. GILBERT'S RECOMMENDATIONS FOR APPLICATION OF STATISTICAL METHODOLOGY IN COMPARING SITE AND BACKGROUND ENVIRONMENTAL DATA

D. R. Weier, A. D. Palachek,
T. R. Gatliffe, D. M. Splett, D. K. Sullivan

Statistical Applications

August 31, 1993

SA/93-017

Statistical Applications

EG&G Rocky Flats Inc.
Rocky Flats Plant
P.O. Box 464
Golden, Colorado 80402-0464

[redacted]

Approved: DR Weier

Executive Summary/Introduction:

Dr. R. O. Gilbert's recommendations for the use of statistical methodology in comparing site and background environmental data are reviewed. Gilbert's recommended approach is reasonable, technically sound, and its application need not negatively impact costs or schedules. It incorporates the nonparametric scores methodology proposed and used by Statistical Applications (SA) personnel; this methodology and supporting computer code were given to Gilbert by SA personnel. His approach is generally quite conservative in that its application will likely minimize the chance of missing site contaminants at the expense of increasing the likelihood of falsely declaring analytes as contaminants when in fact they are not.

Discussion:

The discussion in this report is divided into three sections. The first is the technical review of Gilbert's recommendations; the second is the Statistical Applications proposal for an approach which contains minor modifications of Gilbert's approach; and the third is comments on implementation requirements. Per the request made by D. M. Smith of Environmental Remediation Management, comments are kept somewhat brief.

Note that the particular media of interest may well dictate modification in the sequence of statistical methodology applied. Gilbert's approach is probably most applicable to soils, in particular with respect to the "hot measurement" testing, and this report is written with that application in mind.

Technical Review:

The approach recommended by Gilbert is quite similar to that already used by SA personnel in applications to OU2 and terminal pond data. Differences are that Gilbert recommends some alternative graphical displays, the use of a "hot measurement (HM)" screen, and the addition of two non-parametric tests for elevated site data relative to background data. These two tests are the slippage test and the quantile test which potentially can detect special types of contamination that might be missed by the more common tests for differences of means or medians.

Gilbert recommends five phases:

1. Planning
2. Data collection/validation
3. Data presentation
4. Statistical Tests
5. Professional judgement and geochemical analyses

SA personnel certainly agree with the features of Phases 1, 2, and 5, and they look forward to opportunities to participate in such activities in the future. For this report, comments will be limited to the more technical content of Phases 3 and 4.

For Phase 3 methods, Gilbert recommends ordered listings of data, histograms, boxplots, and probability plotting. These features would all be useful, but the magnitude of generating, analyzing, and presenting all of them would likely be pretty overwhelming considering the many analytes over potentially many areas within an OU which need to be compared to background. The resulting reports would be extremely voluminous. SA personnel feel the spirit and the information of the listings, histograms, and boxplots could be mostly obtained by using multiple boxplots for data with no nondetects and a "graphical ordered listing" for those with nondetects. This would significantly reduce the bulkiness of reports without sacrificing essential information. Examples of these graphical displays will be given in the following section.

SA personnel would prefer to apply formal tests for underlying distributions instead of relying on the subjective interpretation of probability plots. Such plots would be useful in the absence of the capability of performing the formal tests, but not essential when the computing resources are available to perform the formal tests. Note that the method used for replacement of nondetects is very influential on the distributional testing results. Gilbert doesn't discuss this at this point and simply replaces nondetects with the detection limit which may be inappropriate for the distributional testing.

The sequence of proposed statistical criteria in Phase 4 are as follows: hot measurement (HM) comparison, slippage test, quantile test, Wilcoxon rank sum (WRS) test, Gehan test, and t-test.

The Gehan test is simply the nonparametric scores test previously proposed and used by SA personnel in the two SA reports and given to Gilbert while he was in Denver. He often refers to these reports as "Palachek et al." in his recommendations. The WRS test can actually be omitted from the list since with nondetects, Gilbert recommends using the Gehan test instead, and with no nondetects the Gehan test reduces to the WRS test. When the terminology "Gehan test" is used in the remainder of this report, the reader should realize it refers jointly to the Gehan test in the presence of nondetects and the WRS test in the presence of no nondetects.

For each of these statistical tools, except the Gehan test, special treatment needs to be given when nondetects are present in the data. This situation is complicated even more when the nondetects are at multiple detection limits. For distributional testing, HM UTL computations, and for the t-test, data replacement for the nondetects is required. For the UTL computations and the t-test, Gilbert recommends using the Helsel approach, and while SA personnel have used a simpler uniform replacement approach in past applications, they will use the Helsel approach as recommended in any environmental applications in which they are involved.

The slippage and quantile tests are only affected by nondetects involving the larger measurements. Even then, since these are nonparametric procedures and can be computed using only ranks, data replacement is not required; Gilbert suggests using the ranks computed in the Gehan test methodology instead.

For the HM comparison, each value in site data would be compared to a target value to check for exceedances indicating potential high localized levels of contamination (hot spots). This is completely reasonable and quite desirable.

However, SA personnel would recommend that the comparison value is some agreed upon, fixed, possibly risk-based quantity. The use of a 95/95 upper tolerance limit (UTL) is not recommended by SA personnel for two reasons:

1. Such UTL estimators are quite volatile with their behavior depending very heavily on unknown underlying distributions and nondetect replacement approaches. Whether a site value is an exceedance may well be determined more by these features for the background data and the resulting large variability in UTL estimators than on the actual magnitude of the site data.

2. A 95/95 UTL would be exceeded by at least one observation in many cases even for the very same background data which were used to compute the UTL. This would also be the case with additional background data if collected, and more importantly, with site data, even when the site data are not elevated relative to background. This is especially true as sample sizes increase. The resulting false alarm rate of incorrectly identifying locations as hot spots when they are not is thus extremely high.

If a "risk-based standard" approach to specifying hot measurement thresholds is not workable, then the use of a 99/99 UTL would be preferred over the 95/95 UTL so that the false alarm rate is reduced. This, as Gilbert points out, will increase the possibility of failing to identify actual hot spots but SA personnel feel this is warranted to help reduce the otherwise overwhelming false alarm rate of the 95/95 UTL.

Note that the flow chart on page 11.A of Gilbert's recommendations indicates that when at least one measurement exceeds the UTL, the formal statistical tests are bypassed with the next step becoming evaluation through professional judgement and geochemical analyses. SA personnel did not believe that this was Gilbert's intent since such an approach would be technically weak and since it seemed to be in contradiction with his comments in the text that decisions should never be based on a UTL comparison alone. A subsequent August 24 phone conversation between Dr. Gilbert and Dr. D. R. Weier of SA confirmed this. Gilbert has forwarded a revised flow chart to DOE and EPA personnel which indicates that the HM comparisons and the battery of formal statistical tests should be done jointly in all cases. A copy of the revised flowchart is attached to this report.

The intent of the slippage and quantile tests is the detection of special contamination phenomenon for which they have better power than the Gehan test or t-test. Due to the nature of contamination at Rocky Flats, SA personnel expect that only very rarely would the slippage, quantile, and t-test p-values generate a PCOC determination that was not already indicated by the Gehan p-value. However, the application of these tests would be straightforward with minimal additional effort required, so their application is appropriate.

SA personnel have already indicated their belief that the Gehan test is the best approach universally, and it has been used in their previous support efforts for OU2 and the Pond Water Quality IM/IRA.

Gilbert recommends the additional t-test when the underlying data can be taken to be normal (but not for the lognormal case). Again the likelihood of

obtaining normally distributed data from both the background and site data and having the t-test give a different result than the Gehan test is quite small. The contribution of adding the t-test to the battery would likely result in less additional power in contaminant detection than the addition of the slippage and quantile tests. None-the-less, the test would be straightforward and require minimal effort and can be included in the battery.

In summary, the Phase 3 and 4 recommendations of Gilbert are appropriate and the sequence of data presentation methods, HM testing, and formal statistical testing should be followed. A more specific sequence which is a slight modification of Gilbert's is discussed briefly in the next section.

Statistical Applications Proposal:

For data presentation, the ordered listing approach would become rather cumbersome for larger data sets and for several areas to be compared to background. SA personnel have considered several graphical approaches for best displaying the relationship between background and site locations and the information in ordered listings. The "best pictures" for conveying the relationships and information are thought to be multiple boxplots by area when no nondetects are present (radionuclide data) and plots by area indicating detects and non-detect levels when nondetects are present (VOA/SVOA's, total and dissolved metals, and water quality parameters). These have been used in past SA reports and examples are provided on the next two pages.

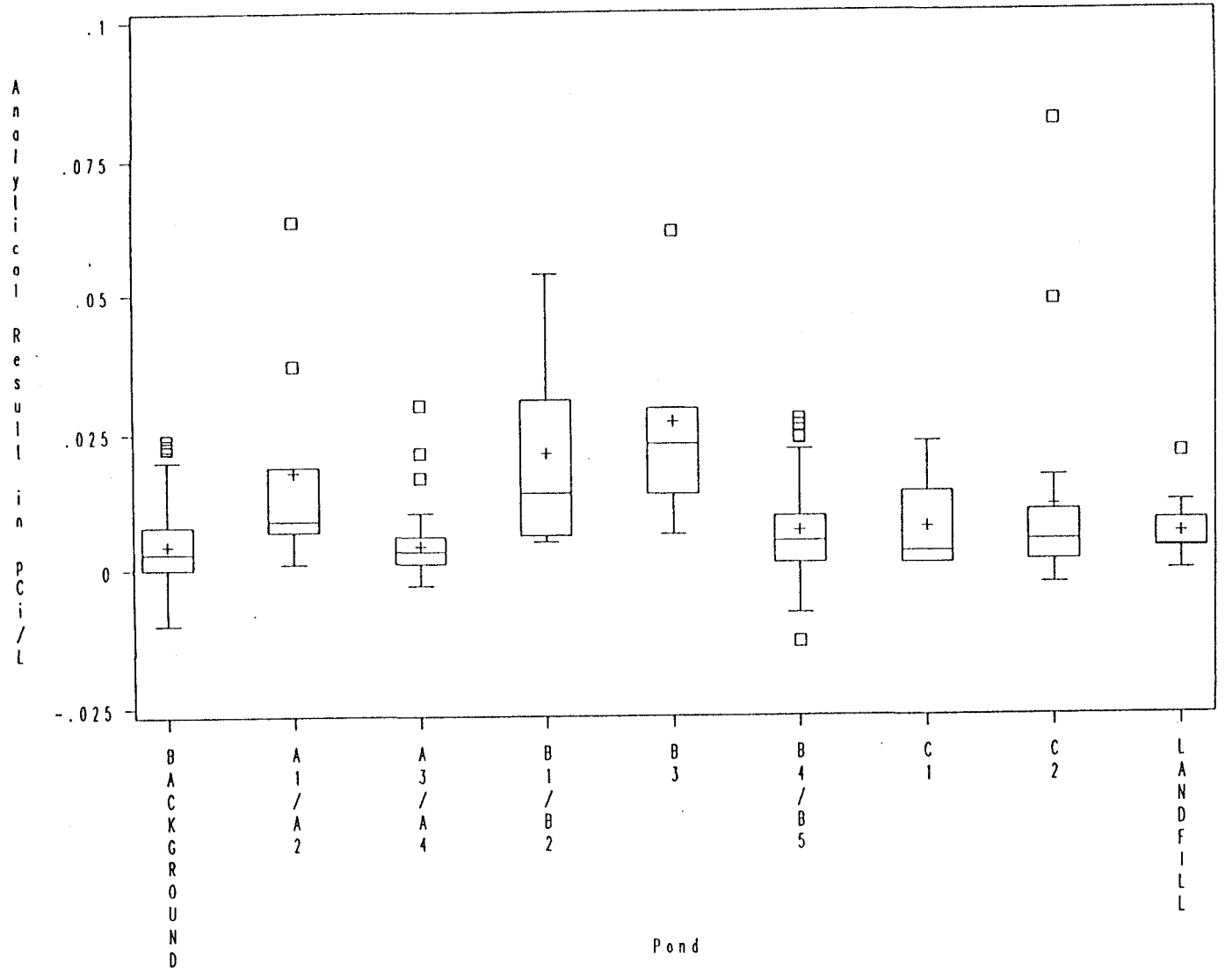
The multiple boxplots on the following page show the relationships between sites and backgrounds although they don't have the detail that the corresponding ordered listings or multiple histograms would. SA personnel believe sufficient information is contained in the multiple boxplots, and their compact nature for many analytes and sites is quite desirable.

In the plot on the second following page, pluses indicate detects and ovals indicate nondetects. These are essentially graphical "ordered listings". Their shortcoming is that each plus and, in particular, each oval can represent multiple measurements. This potential loss of information is again thought to be warranted by the compact presentation of many sites and analytes.

The data presentation would then be followed by the HM and formal statistical testing. If a UTL needs to be computed for HM comparison rather than using a fixed standard of some type, both normal and lognormal results would be presented. P-values for the two distributional tests would be given instead of the probability plotting proposed by Gilbert in the data presentation section. Means, standard deviations, UTL's and other summary statistics as needed, after using the Helsel data replacement approach, would be provided for both the normal and lognormal cases. The larger p-value for the distributional tests would indicate which of the normal or lognormal results is more appropriate. If both p-values are quite small (close to zero), consideration should be given to nonparametric UTL estimates as Gilbert suggests. Note that using the Helsel approach for data replacement for either the normal or lognormal case will dramatically bias distributional tests towards confirming the assumed distribution, especially with substantial numbers of nondetects.

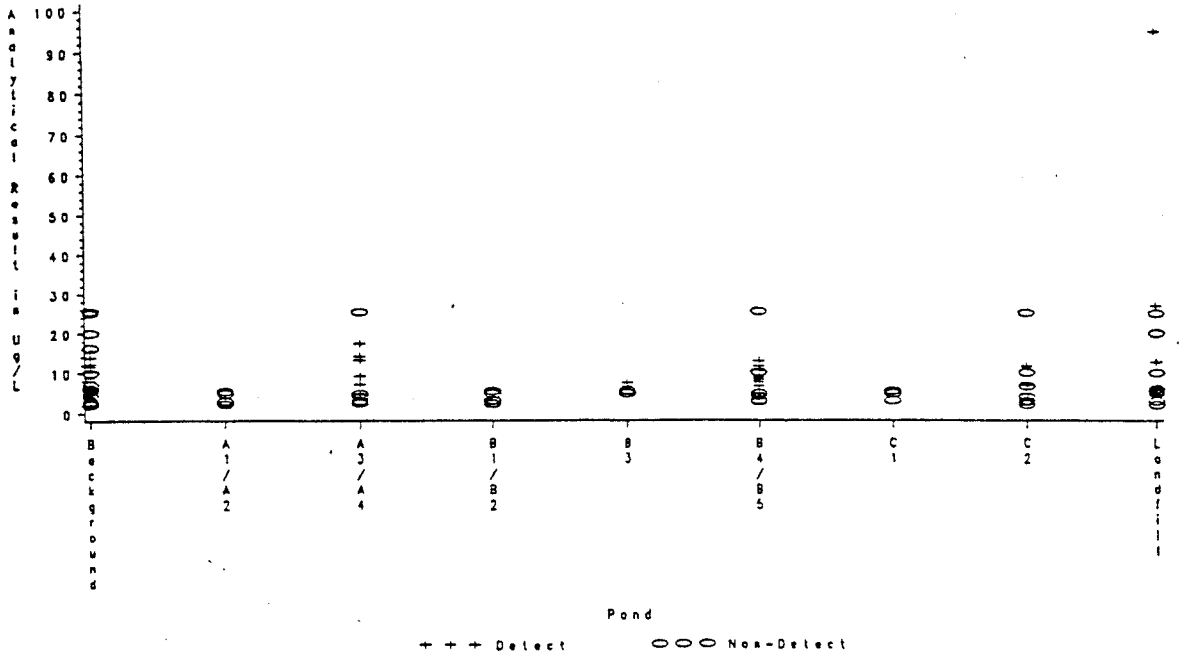
# POND WATER IM/IRA Total Radiochemistry

ANALYTE=AMERICIUM-241



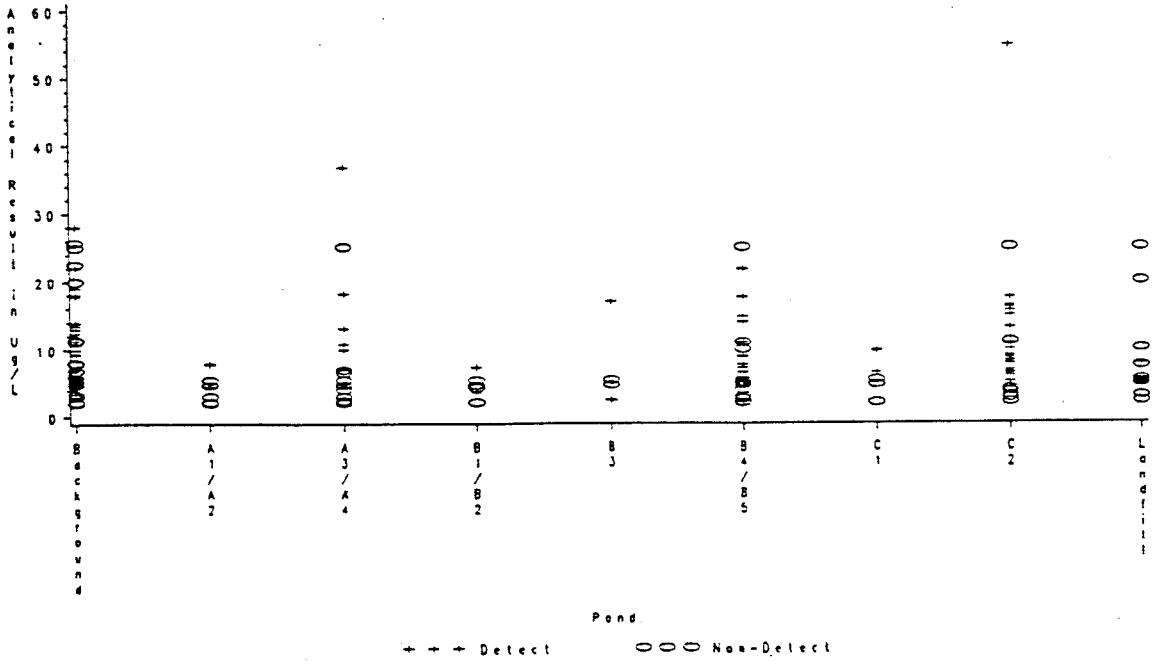
# Pond Water IM/IRA Total Metals

ANALYTE=COPPER




# Pond Water IM/IRA Dissolved Metals

ANALYTE=COPPER



SA personnel will generate computer code for determining p-values for the slippage and quantile tests; these p-values along with the Gehan test p-value (WRS in the case of no nondetects) and t-test p-value would be tabled. Any one of these p-values being sufficiently small would indicate some type of elevated site data in the associated site area. The t-test p-value should only be considered for those cases when normality is found appropriate or for large sample sizes.

Special, unusual cases evident from the data presentation, HM comparisons, and p-value results would be identified and discussed. Final lists of potential contaminants of concern (PCOC) would then be generated for each area in preparation for professional judgement and geochemical evaluation in preparation for risk assessment applications.

Implementation:

The application of the fairly extensive sequence of analysis steps proposed by Gilbert and slightly modified by SA personnel requires substantial computer resources and computing and statistical expertise. Even so, it is fairly straightforward, consisting of little more than the sequence of steps already implemented in previous analyses by Statistical Applications personnel.

As an example, suppose a typical application consists of data from many locations which can be grouped into three areas, and each of these three areas is to be compared to background. The usual set of radionuclides, VOA/SVOA's, total and dissolved metals, and water quality parameters are taken to be of interest. Given background and site data which has already been "cleaned up", with two SA personnel involved, it is estimated that the report containing the proposed SA approach results and conclusions could be generated in about three weeks after receipt of data. If data clean-up is necessary, up to an additional two weeks could be required.

Conclusions:

Gilbert's recommended approach is reasonable, technically sound, and its application need not negatively impact costs or schedules. SA personnel propose an approach which contains minor modifications to Gilbert's recommendations. Required turn-around time for SA personnel to generate such an analysis, conclusions, and associated report is estimated to be about three weeks for clean data with up to an additional two weeks for data which requires initial cleanup.

# TASK 4: FLOW CHART FOR COMPARING OU DATA TO BACKGROUND